

DOCUMENT RESUME

ED 164 606

TM 008 171

AUTHOR Coles, Gary J.
TITLE The Evaluation of the ESEA Title VII Spanish/English Bilingual Education Program: Research Design and Analytic Approaches.
PUB DATE 29 Mar 78
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978); For related document, see ED 154 634
EDRS PRICE MF-\$0.83 HC-\$2.06 Plus Postage.
DESCRIPTORS Academic Achievement; *Bilingual Education; Compensatory Education Programs; *Data Analysis; Elementary Education; English; *Evaluation Methods; *Federal Programs; Language Dominance; National Surveys; Program Effectiveness; *Program Evaluation; *Research Design; Spanish Speaking; Student Attitudes
IDENTIFIERS *Elementary Secondary Education Act Title VII

ABSTRACT

The research design and analytic approaches used in the national evaluation of the Elementary Secondary Education Act (ESEA) Title VII Spanish/English Bilingual Education Program are described. This study evaluated the entire program, as opposed to individual projects, and had four goals: to (1) determine the cognitive and affective impact of bilingual education projects; (2) describe the operating educational processes; (3) identify the educational practices resulting in greater achievement gains; and (4) determine project expenditure per student. The projects selected for data collection were reasonably mature, in either the fourth or fifth year of funding. Title VII students in grades 2-6 and comparable non-Title VII students were assessed in English and Spanish language arts achievement, mathematics achievement, and attitudes. The Comprehensive Tests of Basic Skills (CTBS), the Prueba de Comprension Auditiva Ingles (an English as a second language test), and a Spanish version of the mathematics CTBS were used, depending upon language dominance and proficiency. The statistical methods used are described: analysis of covariance; residual achievement gain analysis; and norm referenced analysis. Evaluation results and policy implications are not contained in this report. (GDC)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

18.21

ED164606

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

The Evaluation of the ESEA Title VII Spanish/English
Bilingual Education Program: Research Design and
Analytic Approaches

Gary J. Coles

American Institutes for Research

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

GARY J. Coles

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM "

TM008 171

Paper presented at the American Educational Research Association meeting
Toronto, Canada, March 29, 1978

Printed in U.S.A.

The Evaluation of the ESEA Title VII Spanish/English
Bilingual Education Program: Research Design and
Analytic Approaches

Gary J. Coles
American Institutes for Research

I. INTRODUCTION

This paper is the first in a three-part presentation describing the recently completed national Evaluation of the ESEA Title VII Spanish/English Bilingual Education Program. As implied by its title, the objectives of this paper are to summarize the Study's research design and analytic approaches. Findings and policy implications are discussed in the papers by Dr. Danoff and Dr. Shoemaker, which follow. More information about the Study and the findings are presented in the following reports.

- Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program, Volume I: Study Design and Interim Findings (AIR Report 48300-2/77-FR[III]), February 1977, (ERIC Report ED 138-090), 490 pp.
- Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program, Volume II: Project Descriptions (AIR Report 48300-2/77-FR[IV], February 1977, (ERIC Report ED 138-091), 599 pp.
- Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program, Volume III: Year Two Impact Data, Educational Process, and In-Depth Analyses (AIR Report 48300-1/78-FR[V], January 1978, 405 pp.
- Evaluation of the Impact of ESEA Title VII Spanish/English Bilingual Education Program: Overview of Study and Findings (AIR Report 48300-3/78-FR[VI], March 1978, 34 pp.

These reports are available from the Office of Planning, Budgeting and Evaluation, U.S. Office of Education, 400 Maryland Ave., S.W., Federal Office Building #6, Room 3044, Washington, D.C. 20202. Volumes I and II are currently in the ERIC system and may be ordered from ERIC Documents Reproduction Service, P.O. Box 190, Arlington, VA 22210. Volumes III and the Overview of Study and Findings should also be available through ERIC in the near future.

II. BACKGROUND AND RATIONALE

A. Program vs. Project Evaluation

The Elementary and Secondary Education Act Title VII Bilingual Education Program was established by Congress in 1968 to meet the special educational needs of students of limited English-speaking ability in the United States so that they could eventually achieve competence in the English language. In 1974, to assess the status of the federally funded bilingual education program in the United States, the U.S. Office of Education's Office of Planning, Budgeting, and Evaluation contracted with the American Institutes for Research (AIR) to conduct an "impact" study of this Program. Up to that time, no large-scale national evaluation of the overall effectiveness of the ESEA Title VII Program had been conducted. While data were collected on individual projects, the goal of the Study was evaluating the impact of the Title VII Program as a whole--not evaluating each individual project. The results reported here describe the Program as implemented during the 1975-76 academic year.

B. Goals of the Study

The goals of the Study were to (1) determine both the cognitive and the affective impact of bilingual education on students in Spanish/English bilingual education projects funded through ESEA Title VII, (2) describe the educational processes operating in these projects, (3) identify those educational practices which result in greater gains in student achievement, and (4) determine per-student costs associated with each project.

C. Projects Selected for the Study

Bilingual education projects selected for evaluation were all Title VII Spanish/English bilingual projects in either their fourth or fifth year of funding as of fall 1975. Projects such as these were thought to be reasonably mature projects--ones having refined their approach to bilingual education over time.

III. DATA COLLECTION STRATEGY

A. Basic Study Design

The general design of the Impact Study was one of contrasting the performance of two groups of students: those enrolled in Title VII Spanish/English bilingual projects and comparable students not enrolled in such projects. Students from grades 2 through 6 in each group were pre-tested and posttested during the 1975-76 school year. A subsample (the Follow-On Sample) of 1975-76 second- and third-grade students was also tested in the fall of the 1976-77 school year. Training, experience, and bilinguality of teachers and teacher aides were determined for both Title VII and Non-Title VII classrooms. In addition, the classroom educational practices in operation in both Title VII and Non-Title VII classrooms were documented thoroughly during the 1975-76 school year for a subsample of all classrooms of students in grades 2 through 6 involved in the Study.

The general design of the Impact Study is illustrated below with the documents associated with each data source indicated.

ANALYSIS SAMPLE	STUDENT TREATMENT AND GRADE										IMPACT STUDY DOCUMENT
	Title VII					Non-Title VII					
	2	3	4	5	6	2	3	4	5	6	
Total Impact Study Sample: Tested Fall 1975 and Spring 1976	X	X	X	X	X	X	X	X	X	X	Volume I Volume III
Intensive Impact Study Sample: Includes all Title VII and comparison sites, but a sample of classrooms and grades were selected for detailed educa- tional treatment documentation during 1975-76 school year	X	X	X	X	X	X	X	X	X	X	Volume II Volume III
Impact Study Follow-On Sample: Sample of 1975-76 school year 2nd and 3rd grades followed into 1976-77 and tested fall 1976											Volume III

B. Sample of Projects, Classrooms, and Students

All projects in their fourth or fifth year of federal funding during the 1975-76 school year were evaluated in the Impact Study. From the total pool of Title VII classrooms present in each of the thirty-eight project sites, a stratified random sample was drawn which included at least one classroom for each site from every grade from second through sixth; moreover, to the extent that participating sites would agree to testing, additional classrooms were randomly chosen so that approximately 40% to 50% of the Title VII classrooms in each participating site were tested. In addition, Non-Title VII classrooms were selected for testing in 20 sites which were able to nominate Non-Title VII classrooms within or near their district whose students were comparable to Title VII students in terms of ethnicity, socioeconomic status, and grade level. In all, over 11,500 students in 384 classrooms in 150 schools at 38 different sites across the United States were initially included in the Impact Study Sample.

The numbers of Title VII and Non-Title VII students in each grade present in both fall 1975 and spring 1976 by ethnic group are given in Table 1. For the Follow-On Sample of second- and third-graders followed into their next school year, there were 191 Title VII second-graders, 63 Non-Title VII second-graders, 201 Title VII third-graders and 81 Non-Title VII third-graders.

C. Instruments and Testing Schedule

Instruments were selected or developed to measure student performance in English Oral Comprehension and Reading, Mathematics, and Spanish Oral Comprehension and Reading. In addition, information was collected via a student questionnaire about student background factors (as, for example, ethnicity, socioeconomic status, sex, and age) and student attitude toward school-related activities. To assess teacher and teacher aide qualifications (for example, training, experience, and bilinguality), a teacher/teacher aide questionnaire was developed. To document the educational experiences of a subsample of students in grades 2 through 6, instruments were designed for structuring interviews conducted with personnel at those sites selected and for observing classrooms.

TABLE 1

Numbers of Title VII and Non-Title VII Students by Ethnic Group
Fall 1975 to Spring 1976 Achievement Analysis, Grades 2, 3, 4, 5, and 6

Grade and Ethnic Group	Number of Students Present During Both Fall 1975 and Spring 1976 Data Collection	
Grade 2	Title VII	Non-Title VII
Hispanic	813	321
White, Non-Hispanic	79	92
Black	74	50
Other	90	33
Grade 3	Title VII	Non-Title VII
Hispanic	1031	520
White, Non-Hispanic	132	125
Black	89	61
Other	118	43
Grade 4	Title VII	Non-Title VII
Hispanic	1025	551
White, Non-Hispanic	115	135
Black	85	71
Other	63	35
Grade 5	Title VII	Non-Title VII
Hispanic	827	154
White, Non-Hispanic	73	60
Black	51	42
Other	107	9
Grade 6	Title VII	Non-Title VII
Hispanic	400	76
White, Non-Hispanic	67	46
Black	20	31
Other	52	5

Each student was assigned a testing package containing tests appropriate to his/her level of language competence in English and Spanish as judged by the classroom teacher. All students were tested in the fall and spring of the 1975-76 school year. The Follow-On second- and third-graders were tested additionally in the fall of the 1976-77 school year.

During site visits made by AIR bilingual-bicultural staff during the 1975-76 school year, the Title VII project director and other project staff were interviewed so that project goals and procedures could be documented. In addition, the educational experiences of students in selected classrooms in the subsample of both Title VII and Non-Title VII classrooms were documented through interviews with teachers and classroom observation.

IV. ANALYSIS VARIABLES AND COVARIATES

A. Measures of Student Achievement and Attitudes Analyzed

In keeping with guidelines received from the U.S. Office of Education, four areas of student achievement were assessed: English Language Arts, Mathematics, Student Attitudes, and Spanish Language Arts. English Language Arts achievement was assessed by the Comprehensive Tests of Basic Skills (CTBS), a widely used achievement test, and the Prueba de Comprension Auditiva Ingles (an English as a Second Language [ESL] test). The ESL test measures understanding of spoken rather than written English and was administered only to those students judged by their teachers to be monolingual Spanish for test-taking purposes. Mathematics achievement was assessed by the CTBS. English and Spanish versions of the Mathematics tests were used. Each student took the test in the language more appropriate for him or her as judged by the classroom teacher. Three student attitudes were measured by a student questionnaire administered in either English or Spanish: Attitude toward English, Attitude toward Spanish, and Attitude toward School. Two Spanish Language Arts measures were analyzed: an orally administered Spanish as a Second Language (SSL) test (given to all students except those judged by their teachers to be monolingual Spanish for test-taking purposes) and the Prueba de Lectura, a test of Spanish Reading (given to all students except those judged to be English monolingual for test-taking purposes by their teachers).

B. Covariates

Pretest Scores

Title VII/Non-Title VII differences in fall 1975 (pretest) scores (along with the other covariates discussed below) were used to adjust post-test differences by means of two different statistical procedures to be presented later.

Student Socioeconomic Status Score

A student Socioeconomic Status score was developed from Student Questionnaire items and provided a measure of the educational and occupational level of the child's parents (or heads of household).

Approximate Number of School Days Intervening Between Pretest and Posttest

An index of the approximate number of school days intervening between the pretesting and posttesting dates for the students tested within each project was developed given known testing dates and opening/closing school dates.

Language Use Indices

Because no suitable group-administered language-dominance assessment instruments were available during the design and data collection phases of the Impact Study, the study could not obtain a test score that would permit a given student to be categorized in terms of his/her "best language for instruction." Nevertheless, because the Title VII and Non-Title VII student groups compared may have differed greatly with respect to use of English and/or Spanish and because such group differences may have somehow "biased" the analyses of Program impact relying on Title VII/Non-Title VII comparisons, project staff implemented a series of analyses that explored whether or not a reliable and valid index of language use could be developed from the data at hand. Such an index or indices, then, could be used as covariates in all subsequent comparative analyses of Program impact and in the analyses relating instructional processes and teacher characteristics to gains in outcomes. The remainder of this section describes the rationale and the procedures used, as well as the results of these index development analyses.

The issues involved in defining and measuring the concepts of "limited English-speaking ability" (LESA) and "best or most appropriate language for instructional purposes" are extremely complex and have been beyond the scope of the Impact Study contract. Thus, it needs to be emphasized that the indices we proposed to develop were not thought of as direct measures of these constructs. Rather, we merely hoped to develop a measure or a series of measures that could be used to rank order, in a meaningful way, the students participating in this study in terms of their language usage.

The first step in these analyses was to identify a series of individual items or multi-item scales present in the Impact Study data base that, on a priori grounds, appeared to measure use of English and/or Spanish. Because such items/scales were judgments, it was desirable that they reflect as many

different kinds of judgments as possible. Table 2 shows the items/scales identified. An examination of the content of these items shows that each asks for an assessment of speaking and understanding, rather than reading and writing, of Spanish or English. Second, the coding scheme shown for the items in which a smiling, frowning, neutral or blank face was marked by the student reflected the fact that students were instructed to mark the blank face if a given item did not apply to them (e.g., speaking English at home). Thus, we assumed that if an opinion toward the use of a language was inappropriate, it was because the student did not speak that language. Lastly, the language in which the student-completed questionnaire was taken (the first item listed in Table 2) seemed appropriate here because it reflected aural comprehension rather than literacy, since all such questionnaires were administered orally.

After identifying and coding the variables shown in Table 2, a matrix of pooled within-groups inter-item correlations was computed. The "group" variable here was Title VII (pooled)/Non-Title VII group membership. This was done so that index development would not be influenced by between-group variation in item responses. This procedure utilized the data only for every other student in the Impact Study data base so that any resulting indices could be cross-validated on data not involved in index development. Correlations were based only on students from all grades with complete data on all eleven items/scales entering the analysis.

The pooled within-groups inter-item correlation matrix for the calibration sample (i.e., odd-numbered students) was then subjected to a principal components analysis. Table 3 shows the loadings for the first two unrotated principal components for the calibration sample (odd-numbered students, $N = 3588$) and for similar analyses based on the hold-out or validation sample (even-numbered students, $N = 3645$) that were implemented later.

Inspection of the pattern of loadings derived in either analysis and of the manner in which the items/scales were coded suggested that Principal Component or Factor I could be named "Preference for English versus Spanish" and that Principal Component or Factor II could be labeled "Use

TABLE 2

Items and Scales Identified in Developing Indices
of English and/or Spanish Use

Item (from fall 1975)	Source	Alternatives (and Coding)*
Language used to complete the Student Questionnaire	Student-Completed Student Questionnaire	English (2) Spanish (1)
The language I speak the best is	Student-Completed Student Questionnaire	English (3) Spanish and English about the same (2) Spanish (1)
I like to speak Spanish	Student-Completed Student Questionnaire	A lot (1) It's O.K. (1) Not much (1) I don't like it at all (1) I don't speak Spanish (2)
I like to speak English	Student-Completed Student Questionnaire	A lot (2) It's O.K. (2) Not much (2) I don't like it at all (2) I don't speak English (1)
At school, when I speak Spanish	Student-Completed Student Questionnaire	Blank face (1) Smiling face (2) Frowning face (2) Neutral face (2)
At school, when I speak English	Student-Completed Student Questionnaire	Blank face (1) Smiling face (2) Frowning face (2) Neutral face (2)
At home, when I speak Spanish with my family, I feel	Student-Completed Student Questionnaire	Blank face (1) Smiling face (2) Frowning face (2) Neutral face (2)
At home, when I speak English with my family, I feel	Student-Completed Student Questionnaire	Blank face (1) Smiling face (2) Frowning face (2) Neutral face (2)

TABLE 2 (continued)

Item (from fall 1975)	Source	Alternatives (and Coding)*
Students' Perceived Proficiency in English (2 item scale):	Student-Completed Student Questionnaire	
Can you understand English?		Yes, very well (3) Yes, more or less (2) No, I can't (1)
Can you speak English?		Yes, very well (3) Yes, more or less (2) No, I can't (1)
Students' Perceived Proficiency in Spanish (2 item scale):		
Can you understand Spanish?	Student-Completed Student Questionnaire	Yes, very well (3) Yes, more or less (2) No, I can't (1)
Can you speak Spanish?		Yes, very well (3) Yes, more or less (2) No, I can't (1)
The child's dominant language is	Teacher-Completed Student Questionnaire	English (2) Spanish (1)

* The coding of items/scales shown here does not necessarily correspond with that used in other portions of the Study.

TABLE 3

Loadings for Eleven Items/Scales on Two Unrotated Principal Components, Calibration, and Validation Samples

Item/Scale	Direction of Coding, English (E) or Spanish (S)	Principal Component Loading			
		I		II	
		Calibration Sample	Validation Sample	Calibration Sample	Validation Sample
Language Used to Complete the Student Questionnaire	E	.60	.62	.21	.29
The Language I Speak Best Is	E	.79	.79	-.03	.00
I Like to Speak Spanish	E	.55	.50	-.39	-.54
I Like to Speak English	E	.25	.23	.50	.63
At School, When I Speak Spanish	S	.00	-.05	.44	.24
At School, When I Speak English	E	-.05	-.06	.43	.16
At Home, When I Speak Spanish with My Family, I Feel	S	-.02	-.05	-.40	-.34
At Home, When I Speak English with My Family, I Feel	E	-.11	-.21	.40	.10
Student's Personal Proficiency in English	E	.57	.57	.45	.51
Student's Personal Proficiency in Spanish	S	-.70	-.68	.39	.45
The Child's Dominant Language	E	.74	.75	.07	-.01
Proportion of Variance Accounted for		24.80	24.60	13.70	13.20

of Both English and Spanish." Factor scores indices were then computed for all calibration sample students by first converting each of the eleven items to a standard score where the mean and standard deviation used was that computed for all calibration sample students with data for a given item (i.e., these calculations were not restricted to students with complete data on all eleven items). Factor Score I, Preference for English versus Spanish, was computed by summing the standard scores for

- Language used to complete the Student Questionnaire (E)
- The language I speak best is (E)
- I like to speak Spanish (E)
- I like to speak English (E)
- Student's Perceived Proficiency in English (E)
- The child's dominant language is (E)

subtracting the standard score for

- Student's Perceived Proficiency in Spanish (S)

and dividing the result by the square root of the number of nonblank item standard scores involved in the computation. Dividing the result by the square root of the number of items added (or subtracted) was done to equate the variance of factor scores based upon sums of differing numbers of items.

Factor Score II, Use of Both English and Spanish, was computed by summing the standard scores for

- I like to speak English (E)
- At school, when I speak Spanish (S)
- At school, when I speak English (E)
- At home, when I speak English with my family I feel (E)
- Student's Perceived Proficiency in English (E)
- Student's Perceived Proficiency in Spanish (S)

and subtracting the standard scores for

- I like to speak Spanish (E)
- At home, when I speak Spanish with my family, I feel (S)

The result of these computations was also divided by the square root of the number of items present.

The procedure of adding and subtracting standard scores to create factor scores in effect weighted each item by 1.0 or -1.0. This was done to simplify the computation of these scores and seemed justifiable because the factor

loadings for each of the items used to construct a given factor score were, in most cases, relatively similar in magnitude.

The validity of Factor Scores I and II was evaluated by examining their relationship with the teacher's assessment of the student's language dominance, a single item on the teacher-completed portion of the Student Questionnaire. Analyses showed that students designated by their teachers as "monolingual Spanish speakers" got low scores on Factor Score I, students designated as "monolingual English speakers" obtained high scores on Factor Score I, and students identified as "bilingual English/Spanish speakers" scored between the two monolingual groups. The same pattern of differences among means on Factor Score I occurred at every grade level. These results suggested that Factor Score I could be viewed as a language use index and that its label "Preference for English versus Spanish" was appropriate.

Mean differences for the calibration sample on Factor Score II were smaller than for Factor Score I. Nevertheless, at every grade level, the means for students designated as "bilingual English/Spanish speakers" was greater than that for either monolingual group. These results, when considered along with those for Factor Score I, suggested that Factor Score II could be viewed as providing a relative measure of a construct called Use of Both English and Spanish.

These conclusions were, of course, dependent upon whether or not the criterion variable used, a single item on which the teacher indicated the student's spoken-language skills, was a credible criterion. Common sense does suggest that teachers should be able to identify accurately whether or not a given student is a monolingual Spanish speaker or a monolingual English speaker. This seems especially reasonable for the Title VII and Non-Title VII teachers sampled by the Impact Study because (1) many teachers claimed to be Spanish/English bilingual or, at least, to have spoken Spanish sometime during their lifetime; (2) the large numbers of Hispanic students in both Title VII and Non-Title VII classrooms or schools would mean that most teachers would be exposed to Spanish-speaking children and should, at the very least, be able to identify Spanish when they heard it; and (3) teachers had sufficient time to observe their students in a number of different academic and social situations since the

teacher-completed student questionnaire was administered two or more months after the beginning of school in fall 1975.

For these reasons, it seems likely that the designations "monolingual English speaker" and "monolingual Spanish speaker" were accurate. This being the case, what can be said about the accuracy of the designation "bilingual English/Spanish speaker"? The most logical answer is that a student designated as a "bilingual English/Spanish speaker" was not a Spanish or English monolingual speaker. Whether or not the students given this designation were truly bilingual cannot be inferred. However, the facts that the factor score indices were created from meaningful combinations of items concerning language use and that expected group differences on factors were observed do tend to validate both the factor score indices and the credibility of the item used as the criterion variable in these analyses.

To ascertain if these results were stable and not due to capitalizing on chance relationships among items in the calibration sample data, similar computations were performed on validation sample students. As shown in Table 3, the principal component solutions for the calibration and validation samples were very similar. Table 4 shows the same pattern of results for calibration and verification sample students even though the factor scores for validation sample students were computed using the scaling rules and the means and standard deviations derived in analyses of calibration sample data.

These results provide evidence that the factor score indices, Preference for English versus Spanish and Use of Both English and Spanish, can be considered valid and reliable measures of their respective constructs. Factor scores were then computed for each student in the Impact Study data base who had at least one of the items entering an index by computing overall means and standard deviations, by computing standard scores, and by summing (or subtracting) item standard scores in the manner described previously.

Since a given factor score could be computed for a student if at least one item contributing to the score was present the amount of missing data was minimal (11% and 17% for Factor Scores I and II, respectively). Nevertheless, it was desirable to include in analyses all students with pretest and posttest scores and to avoid having to delete students because of

TABLE 4

Calibration Sample and Validation Sample Means and Standard Deviation
for Factor Score 1 and Factor Score 2 by Teacher Assessment of Student
Language Dominance--All Students

Teacher Assessment of Student Language Dominance	Factor Score 1		Factor Score 2	
	Calibration Sample	Validation Sample	Calibration Sample	Validation Sample
Monolingual Spanish Speaker				
Mean	-1.93	-1.95	-0.14	-0.23
Standard Deviation	1.17	1.21	1.20	1.22
N	564	567	518	515
Bilingual English/Spanish Speaker				
Mean	-0.31	-0.29	0.26	0.27
Standard Deviation	1.08	1.04	1.16	1.07
N	2727	2721	2555	2561
Monolingual English Speaker				
Mean	1.77	1.82	-0.60	-0.58
Standard Deviation	1.00	0.94	1.24	1.23
N	1110	1102	1023	1025
F	2520.71	2788.30	197.24	218.43
n ²	0.53	0.56	0.09	0.10

missing data on Factor Scores I and II. Thus, missing data was replaced for students who were present during fall 1975 and spring 1976 with the mean score obtained by all students who had data, who were of the same ethnic group (Hispanic, white non-Hispanic, black, and other), who were also present during both fall 1975 and spring 1976, and who were members of the same treatment group (Title VII With Comparison, Title VII Without Comparisons, and Non-Title VII). Because Hispanic students represent the target population of this particular study, grade level was also taken into account in any missing data replacement for such students.

Lastly, because Factor Scores I and II were, by design, not correlated with each other, it was desirable to create a third factor score variable to be used as a covariate along with Factor Score I (Preference for English versus Spanish) and Factor Score II (Use of Both English and Spanish) in later impact analyses. A product variable was generated which, when used along with Factor Scores I and II, would permit statistical adjustments for both factor constructs as well as their interaction. These three variables, then, would permit Title VII/Non-Title VII comparisons and the process analyses to be adjusted for systematic student differences with respect to

- High Preference for English versus Spanish and
High Use of Both English and Spanish
- High Preference for English versus Spanish and
Low Use of Both English and Spanish
- Low Preference for English versus Spanish and
High Use of Both English and Spanish
- Low Preference for English versus Spanish and
Low Use of Both English and Spanish

V. ANALYSIS ISSUES AND STRATEGIES

A. Revised Evaluation Design

The basic design employed by the Impact Study was, as described previously, a two-group pretest/posttest design in which the growth of students enrolled in Title VII-funded classrooms was compared with that of similar students enrolled in Non-Title VII, non-bilingual education classrooms. Unfortunately, because of community practices regarding the student composition of Title VII-supported classrooms and because of legal constraints involving segregation, Non-Title VII, non-bilingual education classrooms containing students similar to those served by Title VII could not always be identified at each site. Since students enrolled in such Non-Title VII schools/classrooms were used to form the comparison group in most impact analyses, it became necessary to ascertain if there were notable systematic differences among sites that had or had not identified comparison schools or classrooms. If comparison classrooms were located in a clearly definable subset of sites, then it was possible that the growth of such Non-Title VII students might not form a completely appropriate "benchmark" group against which to evaluate the growth of all Title VII students. Our general analysis approach, then, was formulated in a way that would permit us, if necessary, to take such differences into account and would allow us

- to conduct appropriate Title VII/Non-Title VII comparative analyses
- to identify the assumptions that were necessary in making program-level inferences from a given set of comparative analyses

The revised evaluation design can be described by the following figure in which classrooms and students had been sampled in only three of the four cells. Since community type is defined by whether or not "matched" Title VII and Non-Title VII classrooms were sampled, the Non-Title VII Type WOC cell was, by definition, blank.

Community Type

<u>Treatment</u>	With Comparison (Type WC)	Without Comparison (Type WOC)
Title VII	Cell 1	Cell 2
Non-Title VII	Cell 3	

Within this design, it is clearly possible to compare the outcome of Title VII treatment to what it might have been without treatment, within Type WC (with comparison) communities. Overall comparison involving Title VII classrooms, from both community types, however, depends on additional assumptions. First, if we make the assumption of no interaction (the effect of Title VII treatment would be the same in both types of communities), then a Title VII/Non-Title VII comparison within Type WC communities would suffice as an overall Title VII/Non-Title VII comparison. Second, we can make the (weaker) assumption that, if there are not significant differences in the effectiveness of the Title VII project treatments between communities of Types WC and WOC, there are not likely to be differences in the effectiveness of Non-Title VII project treatments between communities of Types WC and WOC. The rationale for this assumption is that any differential effectiveness of Title VII treatments between community types would be mediated by some community variable which would also have some significant effect--but maybe of a different size--on the effectiveness of the Non-Title VII treatment. Given this assumption, if community type had no effect on Title VII outcomes, Title VII scores could be pooled and compared with Non-Title VII scores.

Neither of these assumptions provides for a truly adequate substitute for comparison schools from both types of communities. However, each assumption does allow for a useful assessment of Title VII effectiveness. To have obtained comparison schools in all sites, it would have been necessary to substantially alter the program design in some communities, an endeavor far beyond the mandate for this study.

In certain cases where Title VII treatments had different effects in Type WC and WOC communities, a further comparison was made between two groups--Title VII classrooms in Type WOC communities versus Non-Title VII classrooms in Type WC communities. This was primarily for descriptive purposes. For example, suppose that in Type WC communities Title VII and Non-Title VII treatments were equally effective but that the positive impact of Title VII treatment in Type WOC communities was substantial. Then one could reasonably entertain the possibility that, had comparison classrooms been available in all communities, Title VII might have been shown effective overall. If, on the other hand, the Title VII treatments in Type WOC communities were significantly less effective than either Title VII or Non-Title VII treatments in Type WC communities, it is more doubtful that the universal availability of comparison classrooms would have improved the overall standing of Title VII versus Non-Title VII classrooms, compared to the results obtained in Type WC communities alone.

Preliminary analyses, in fact, showed that presence or absence of an identified Non-Title VII classroom was confounded with geographic area and Hispanic subgroup. Title VII classrooms with identified Non-Title VII comparison classrooms (community Type WC) were usually located in the western or southwestern region of the United States and were composed of students whose Hispanic background was Mexican-American. On the other hand, Title VII classrooms with no identified Non-Title VII comparison classrooms (community Type WOC) were mainly located in the eastern or northeastern region of the United States and were largely composed of students whose Hispanic background was Puerto Rican-American. This being the case, the community type variable could not be ignored.

B. Analytic Approaches

A decade of experience in the evaluation of federal education programs has shown that no single method of statistical analysis can be counted on to permit the evaluator to reach completely valid and uncontestable conclusions. This is because statistical analytic methods are developed to apply to given idealized situations and because the organization and implementation of educational programs are only rough approximations to the idealized situation upon which any one method is based. One solution to

this problem is to (1) analyze the data using several kinds of statistical methods, (2) examine empirically (if possible) the degree to which the assumptions underlying each method have been met, (3) look for converging lines of evidence across methods, and (4) limit conclusions to those findings that appear plausible across methods. Thus, three different methods were used to assess, across grades, Program impact on gains in student achievement: (1) standardized residual gain score analysis, (2) analysis of covariance, and (3) comparison of pretest and posttest percentile ranks. In all analyses concerning Title VII impact, the question being explored was whether or not the growth in achievement and attitudes for Title VII Hispanic students were greater than that expected in the absence of a special Title VII educational treatment. (Most analyses were limited to Hispanic students because they included the target population of interest to Spanish/English projects.)

Analysis of Covariance (ANCOVA)

There are no theoretically adequate methods for comparing the performance of groups formed through nonrandom processes. The only exception occurs when all factors involved in assigning students to groups are known--a highly unlikely state of affairs in field studies. One of the most popular solutions to this difficulty in applied research undertaken in public schools is to ascertain whether the student composition of the groups to be compared is relatively similar so that it is reasonable to assume that group membership is random in effect (Lord, 1967). This approach, for example, underlies the comparison group model currently proposed for use in evaluating Title I projects (Tallmadge and Wood, 1976, p. 50):

For assignment to be random in effect, it must be reasonable to assume that the treatment and control groups are, in essence, random samples from a single population. To be considered random in effect, both groups of children must be similar in terms of socioeconomic status, age, sex, and racial composition, and they must attend schools which are similar in size and setting.

Furthermore, the method of selecting comparison group students should be similar to the method used to select Title VII students to avoid differential

"regression-to-the mean" effects in the two groups. Perhaps the most important characteristic to consider is students' pretest scores, since group differences on the posttest are used to evaluate Program impact.

When pretreatment differences among groups to be compared are small or unsystematic so that it is reasonable to assume that the two groups can be considered to be random samples from the same population (prior to the treatment), then analysis of covariance (ANCOVA) can be used to adjust posttest differences for known group differences existing prior to exposure to the treatment.

It is important to note that ANCOVA and standardized residual gain score analysis (described next) assume that the pretest measure is actually a measure of pretreatment achievement level. Because prior participation in Title VII may have had some impact on achievement and because pretest scores may reflect this, the analyses conducted for the Impact Study cannot be viewed as providing evidence for the impact of Title VII during previous school years. Rather, the treatment of interest in this study was the impact during the interval between pretesting and posttesting.

Standardized Residual Gain Score Analysis

Many researchers have shown that when two groups compared with each other cannot be considered to be random samples from the same population (prior to treatment and pretesting), ANCOVA will undercorrect for systematic pre-existing test score differences among the groups. This argument, of course, applies to the situation in which the special treatment is given to the initially low scoring group, an expected occurrence in evaluations of compensatory education programs. Because so many authors have discussed this particular problem in using ANCOVA, there seems little doubt that such regression artifacts can occur. O'Connor (1973), as well as others, have also shown that "matching" suffers from the same kind of regression artifact.

Because of the possibility that the Title VII and Non-Title VII groups to be compared could not be considered to be, even on a practical level, random samples from the same population (or from populations that behave similarly), an analytic alternative to ANCOVA was desired.

Standardized gain score analysis, a procedure described fully by Kenny (1975), was selected as such an alternative analytic approach. It

also permitted us to adjust gains for the same covariates that were used in the ANCOVA analyses. Kenny has shown that if pre-existing differences (i.e., mean differences between Title VII and Non-Title VII groups on the pretest) can be assumed to remain constant over time, group mean differences on the posttest can be adjusted appropriately by use of standardized gain score analysis. He advocates standardized gain score analysis because separate standardization of all pretest scores (Title VII and Non-Title VII alike) and all posttest scores will equate the variances of these two distributions of scores. Since it can be shown that correlations are indices of the degree of relationship between two pairs of standard scores, he shows that standardized gain score analysis can be implemented by comparing the correlation of pretest and treatment group membership (coded as a binary variable) with the correlation of posttest with treatment group membership. Because treatment is a binary coded variable, such correlations become point biserial correlations and are, in fact, an index of the mean difference (in standard deviation units) among the two groups on the continuous variable (in this case, pretest or posttest). Kenny also describes how such an approach can take into account additional background variables or covariates. When this is done, the analyses, in effect, become standardized residual gain score analyses.

As suggested above, the standardized residual gain score analyses were implemented by means of a computational procedure described by Kenny (1975). He suggests that the background variables to be used as covariates should be chosen on a priori grounds and that "partial correlations of the treatment with the pretest and the posttest, with the background variables partialled out" (p. 353) should be compared. "Partial correlations are compared instead of partial regression coefficients because of the assumption of stability of variance over time; that is, the variability of the dependent variable that is not explained by the background variables is stable over time" (p. 353). Thus, we computed two partial correlations, one between treatment group membership and pretest (with various covariates partialled out) and one between treatment group membership and posttest (with the same covariates partialled out). These were compared with each other by means of a t test for correlated correlations.

Norm-Referenced Analysis

The third analytic method used to evaluate Title VII impact involved transforming each CTBS test score group mean to a percentile rank (using published norms) and comparing the posttest percentile rank with that for the pretest. In spite of the fact that caution needs to be exercised when using norms that have been derived by interpolation (such as CTBS norms for fall testing), it was felt that an analytic approach that was not dependent upon comparisons with the Study's Non-Title VII students would be advisable. Furthermore, since the Follow-On analyses involved fall-to-fall comparisons, it was felt that any biasing factors due to use of interpolated fall norms would probably not differentially affect the estimate of growth.

C. Analytic Issues Related to Interpreting the Study's Results

Several issues that are frequently involved in the analyses of data collected by evaluations of compensatory education programs and that may have affected the validity of the Impact Study's conclusions regarding Title VII impact were examined. These issues and their results are summarized in the following paragraphs.

Attrition and Missing Data

Although there were some significant differences across treatment groups in attrition and test omission effects, their effects on the Program impact analyses were generally not substantial. In the few cases where they might be substantial, employing the estimated corrections would have had the effect of making Title VII gains appear smaller. Thus, differential attrition and test omissions could not have been the cause for any lack of finding significant benefits from Title VII treatment in terms of gains on CTBS scales.

Time of Testing

All the differences between testing times slightly favored the Title VII students, and, if the effects were substantial, they could cast doubt on results showing positive Title VII effects.

Packet Assignment Validity

In the Study's Volume I report, statements were made concerning the proportions of students in Title VII who took one of four test packets based on teacher judgment related to appropriate language for test-taking purposes. Because reviewers of the report questioned the validity of those teacher judgments, packet assignments were compared to the two derived language preference factors. If the teacher packet assignments were valid, then the students assigned the English monolingual packet should have been high on the English Preference factor and low on the Bilinguality factor, English-dominant bilingual packet students should have been relatively high on both factors, Spanish-dominant bilingual packet students should have been low on the English Preference factor but high on the Bilinguality factor, and the Spanish monolingual packet students should have been relatively low on both factors.

The separation of factor scores in the predicted direction between packet-assignment groups was significant and suggests that teacher judgments related to packet assignments and language dominance for test-taking purposes were quite valid.

Prior Bilingual Education Experience

The importance of years of prior bilingual education experience as a modifier of Title VII Program impact during fall to spring of the 1975-76 school year was analyzed by means of multiple regression analysis. A regression model was created that included as predictors of posttest: (1) pretest, (2) years of prior bilingual education experience, and (3) the pretest and years of prior bilingual education experience interaction. The proportion of posttest variance accounted for by the model was then compared with that accounted for by a given regression model that did not include the interaction predictor variable.

In grades 2, 3, and 4, in no case did the years of prior experience add as much as 1% to the posttest variance accounted for. In grade 5, there was a contribution of 4% to the variance in ESL scores, 2% to the SSL scores, and 1% to the Reading Total scores; and in grade 6 there was a contribution of over 1% to the variance only in Prueba de Lectura, ESL, Mathematics Computation, and Attitude Toward Spanish scores. These

exploratory analyses thus showed that years of prior experience in bilingual education accounted for typically no more than 1% of the variance in any of the student outcome measures in English, Mathematics, Spanish, and Attitudes.

Number of Boys and Girls in Each Treatment

An examination of the proportions of males and females by grades within Title VII and Non-Title VII groups revealed an essentially even distribution of males and females within grades and within treatment classification. It seems reasonable, therefore, to assume that sex differences were not associated with treatment classification, and thus no further analyses by sex appeared necessary to assess overall Title VII impact.

Additional Issues

Other issues such as the use of CTBS interpolated norms in the norm-referenced analyses and Title VII/Non-Title VII group differences with respect to differential rates of guessing on test items, test floor and ceiling effects and out of level testing effects were also examined. Results showed that caution would be appropriate in using CTBS norms in assessing fall-to-spring growth in achievement, and that the other issues would not, in general, tend to bias comparative analyses in favor of the Non-Title VII/Comparison group.

Conclusions

It is, of course, only an approximation to try to aggregate the implications of all the findings derived in the analyses of these methodological issues because the joint influences of such issues were not analyzed (e.g., it is impossible to measure guessing rates of individuals not present for testing). Also, most of the methodological analyses pertained only to the CTBS tests, so little can be said about the other tests and measures. In order to obtain a rough estimate of the effect of the distortions on some of the results, we added together the rough "correction" estimates for each and compared them to the estimates of residual posttest mean differences between Title VII With Comparison classrooms and Non-Title VII classrooms shown in the ANCOVA analyses. The only case in which the summed corrections reduced a significant difference to less than one-half its

reported size was for the Mathematics gains in fifth grade. The attrition and test-omission corrections each would reduce the positive Title VII impact by nine Mathematics Expanded Standard Score points, which would reduce the adjusted posttest mean difference from 22 points to 4.

It thus would appear that the comparative analyses upon which conclusions regarding Title VII Program impact were based were, in general, not biased to any significant degree. In fact, if any trend was apparent, it was that the methodological issues discussed may have enhanced positive estimated Title VII Program effects slightly. The following two papers, respectively, will present the Study's findings and their policy implications.

REFERENCES

- Kenny, David A. Quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 1975, 82(3), 345-362.
- Lord, F. M. A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, 68, 304-305.
- O'Connor, Jr., E. F. Why matching analyses of covariance, and multiple regression generally fail to control for pre-existing differences. Paper presented at Western Psychological Association. Anaheim, California, May 1973.
- Tallmadge, G. K., & Wood, C. T. User's guide. ESEA Title I evaluation and reporting system. Mountain View, California: RMC Research Corporation, 1976.